DOCUMENT RESUME

ED 339 745 TM 017 677

AUTHOR Parshall, Cynthia G.; And Others
TITLE Statistical Error in Linear Equating with Small Samples of Examinees.
PUB DATE Nov 91
NOTE 46p.; Paper presented at the Annual Meeting of the Florida Educational Research Association (Clearwater, FL, November 13-16, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Simulation; Elementary School Teachers; Elementary Secondary Education; Equated Scores; *Error of Measurement; Higher Education; Licensing Examinations (Professions); Mathematical Models; Monte Carlo Methods; *Sample Size; Secondary School Teachers; *Statistical Bias; Teacher Certification; Test Bias; *Testing Problems
IDENTIFIERS *Linear Equating Method; Nonequivalent Control Groups; Subject Content Knowledge

ABSTRACT

A Monte Carlo study was conducted to compare the statistical bias and standard errors of non-equivalent-groups linear test equating in small samples of examinees. One thousand samples of each size (15, 25, 50, and 100) were drawn with replacement from each of five archival data files from elementary school and secondary school teacher subject area tests (in the areas of early childhood education, art education, music education, mathematics, and Spanish). Equatings were also conducted on 100 samples of size 500 to provide a basis for the interpretation of the standard errors in the small samples. For each test, data files from two parallel forms were used. Results suggest trivial levels of equating bias even with small samples, but substantial increases in standard errors as the sample size decreases. Results are interpreted in terms of applications to testing situations in which small numbers of examinees are available. Four tables and 10 graphs present study data. A 13-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
CYNTHIA G. PARSHALL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED339745

Statistical Error in Linear Equating with Small Samples of Examinees

Cynthia G. Parshall
Department of Educational Measurement and Research
University of South Florida

Pansy Du Bose
Institute for Instructional Research and Practice
University of South Florida

Jeffrey D. Kromrey
Department of Educational Measurement and Research
University of South Florida

Paper presented at the Annual Meeting of the Florida Educational Research Association, November 13-16, 1991, Clearwater Beach, Florida

TM017677

BEST COPY AVAILABLE

Abstract

A Monte Carlo study was conducted to compare the statistical bias and standard errors of nonequivalent-groups linear test equating in small samples of examinees. Sample sizes of 15, 25, 50, and 100 were examined. One thousand samples of each size were drawn with replacement from each of five archival data files from teacher subject area tests. For each test, data files from two parallel forms were used. Results suggest trivial levels of equating bias even with small samples, but substantial increases in standard errors as sample size decreases. Results were interpreted in terms of applications to testing situations in which small numbers of examinees are available.

Introduction

Testing programs typically use multiple, parallel forms of a single test, to provide increased test security and to avoid undesirable practice effects. Such test forms are usually statistically equated, to allow scores from any of the individual forms to be placed on a common scale. Many procedures for the collection and analysis of data to provide test equating are available, but the common-item non-equivalent groups design (Angoff's Model IV, 1984) is probably the most often used.

The applicability of test equating to testing programs in which the population of examinees is small is largely unexplored in the literature. In an effort to add to the knowledge base in this area, an empirical investigation was conducted to examine the effects of sample size on the stability and bias of linear equating, using the common-items non-equivalent groups design.

Literature

The appropriate method of test equating to be used in a particular situation is determined in large part by specific aspects of the test construction and administration. One typical test equating design is for two forms of a test to be constructed with a set of common items; the two test forms are then administered to two groups of examinees on different test dates. This design is referred to as the common-item, non-equivalent

groups design or the anchor test design (Design IV, in Angoff, 1984). Since this design does not necessitate the double administrations of a single form of the test or having multiple forms in use at a given administration, it is often the most practical solution. A primary equating method used in the common-item design is linear equating. In linear equating, the scale of one test form is adjusted so that the means and standard deviations of the two test forms are equal to each other.

The study of test equating and its applications is relatively recent development in educational measurement and knowledge of many aspects of equating is still incomplete. Important issues currently being investigated in the common-item design include the adequacy of equating with small sample sizes, the determination of a suitable equating criterion, and the impact of the relationship between the anchor test and total tests on the adequacy of equatings.

Small Sample Size

An important research topic is that of the effect on equating when small numbers of examinees complete each test form. Most of the equating studies reported in the literature have been conducted on samples of several thousand examinees, with very few studies using samples of only a few hundred examinees. The critical aspect of equating with very small samples is that equating errors may become quite large. In addition, the

formulas for estimating the standard errors of equating may not hold under these small sample conditions. That is, the estimates of the precision of the equating are inaccurate. However, the need for test equating does not become unimportant in specialized programs in which the number of examinees is small.

In one of the few studies in which small sample sizes were used, Kolen (1985) examined samples of size 100 and 250 in order to determine the accuracy of estimates of standard errors of equating. He found standard errors of equating, when derived without the normality assumption, to be sufficiently accurate with sample sizes of 250.

Anchor Test Characteristics

Several studies of the common-item design have been concerned with examining critical aspects of the anchor test. The research suggests that the accuracy of test equating can be affected by the length of the anchor test, by the extent to which the anchor test mirrors the total test in content representation and in statistical characteristics, and by the strength of the correlation between scores on the anchor test and scores on the total test.

Budescu (1985) examined the relationship between the length of the anchor test and the efficiency of linear equating. His results indicated that the most efficient equating is obtained

when one-half of the total test consists of anchor test items. Additionally, his results suggest that the magnitude of the correlation between the anchor test and each form of the total test is the single most important determinant of equating efficiency. Klein & Kolen (1985, cited in Cook & Petersen, 1987) also examined the effects of anchor test length on equating accuracy, using linear equating with non-random groups. All of the anchor tests used were comparable to the total test in terms of content representativeness, difficulty, and discrimination; the anchor tests varied in length and consisted of 20, 40, 60, 80 and 100 items. The results indicated that when examinee groups differed in ability, lengthening the anchor test substantially improved the quality of equating.

Klein and Jarjoura (1985), on the other hand, studied the relationship between equating accuracy and the content of the anchor test. They compared content representative anchor tests, in which the percentage of items in each of several content areas matched that of the total test, with nonrepresentative but considerably longer anchor tests. The results indicated that the content representation of the anchor test is more important that the length of the anchor test in determining the quality of equating.

Peterson, Marco and Stewart (1982), in a comprehensive examination of the adequacy of linear equating methods, performed

equatings under 11 different conditions. The variations included differences in test form difficulties and lengths. The anchor test was varied in location, content similarity to the test forms, and difficulty compared to the test forms. The results suggested that an anchor test which is constructed as a miniature of the total test gives the best results. Differences in difficulty, either between the two test forms or between the anchor and the total test, tended to produce a nonlinear relationship, which caused linear equating models to perform poorly.

Indices of Equating Accuracy

A third field of research that has direct practical implications for test equating is that of evaluative indices for the accuracy of equating. Several studies have investigated the use of potential indices as criteria for determining the adequacy of an equating solution.

There is no clear agreement about an appropriate criterion to be used in determining the adequacy of an equating procedure, rather several different methods have been applied. Budescu (1985) recommended using some monotonic function of the correlation between the anchor and the unique components of each form as a measure of the efficiency of linear equating. Kolen and Whitney (1982) used a cross-validation criterion in their comparison of four equating methods and found it to be a useful external index.

Klein and Jarjoura (1985) equated a test to itself through an equating chain. The adequacy of the equating was then determined through an examination of the root-mean-squared error of equating through the chain and the mean equating error or bias. In Angoff & Cowell's (1986) examination of the assumption of population independence, empirical functions for the standard error of equating were developed from the data for the base population. These functions were then used to evaluate the departure of conversions which were obtained from each of the 11 specially defined subpopulations.

Jaeger (1980) studied five statistical indices for determining whether linear equating would be appropriate in a particular situation. The indices he examined were: the similarity of the two cumulative score distributions; the shape of the raw-score to scaled-score transformation; the consistency of linear and equipercentile equating results; the similarity of item difficulty distributions; and the similarity of item discrimination distributions. All of these indices (except the similarity of item discrimination distributions), provided clear distinctions between those situations in which the linear equating method was adequate and those in which it was not. The similarity of item difficulty distributions appeared to be a particularly useful index because it is a purely statistical criterion for evaluating the adequacy of linear equating and

because it can be computed from data collected during test development.

Kolen (1985) examined standard errors of Tucker equating derived with and without the normality assumption as indices of equating error which is due to examinee sampling. Standard errors obtained through Efron's bootstrap method were found to compare closely to these formulaic standard errors in two Monte Carlo studies.

The purpose of the research reported in this paper was to empirically estimate the standard errors of equating and statistical bias in equatings with small samples of examinees. This Monte Carlo study was conducted by drawing small random samples from existing archival test data files, calculating the equating function derived from each sample, and evaluating these functions for statistical bias and sampling error.

Method

Pseudo-populations Examined

The data on which this research was conducted were random samples drawn with replacement from existing archival test data files. These files were treated as pseudo-populations from which samples were drawn. Data files from five state teacher certification subject area tests were used: Early Childhood Education (K-3), Art Education (K-12), Music Education (K-12),

Mathematics (9-12), and Spanish (K-12). For each test, data files from two parallel forms were used. Descriptive data on these pseudo-populations are provided in Table 1.

Insert Table 1 about here

The use of actual test data files rather than computer-generated data provides a more realistic foundation for the research. Actual field data are often only roughly approximated by computer simulation models. The replication of the study over five different examination data files provides evidence of the generalizability of the results.

Sampling and Equating

Random samples of size 15, 25, 50, and 100 were drawn with replacement from the Form A and Form B pseudo-populations in each examination subject area. For each pair of samples (Form A and Form B) Angoff's Model IV linear equating statistics were calculated. One thousand such equatings were computed for each sample size and each subject area examination. To provide a basis for the interpretation of the standard errors in the small samples, equatings were also conducted on 1000 samples of size 500. Samples of this size are clearly within the range of adequate sizes for linear equating.

The adequacy of the equatings were evaluated by calculating the bootstrap standard errors of equating and the statistical bias of the equating. The standard errors of equating are the standard deviations of the obtained equated scores in the bootstrap samples. The standard errors are given by

$$SE(\theta_i) = \sqrt{\frac{\Sigma(\theta_{ij} - \theta_{i\cdot})^2}{n - 1}}$$

where $SE(\theta_i)$ = standard error for equated score i,

θ_{ij} = obtained equated score i in sample j, and

$\theta_{i\cdot}$ = mean equated score i over the 1000 samples.

The statistical bias in the equating was calculated as the difference between the mean equated scores in the bootstrap samples and the population equated scores calculated directly from the pseudo-population data files investigated. The bias is given by

$$Bias(\theta_i) = \theta_{i\cdot} - \theta_i$$

where $Bias(\theta_i)$ = statistical bias for equated score i,

$\theta_{i\cdot}$ = mean equated score i over the 1000 samples, and

θ_i = population equated score i.

Results and Discussion

The obtained estimates of the standard errors of equating for each sample size in each of the five subject-area examinations

are presented in Figures 1 through 5. These figures are graphical representations of the standard errors of equating at all possible raw score points on each test form. The most striking general characteristics of these empirical estimates of the standard errors are not surprising. First, the sampling error is smallest in the proximity of the mean raw score for the examination, and the error increases monotonically (but not linearly) as a function of the deviation of scores from the mean. Second, the standard error at all scores becomes progressively larger as sample size decreases.

Insert Figures 1 - 5 about here

Notable in all of these figures is the tendency for the error curve to rise more sharply with smaller samples of examinees. The implication of this effect is that the differences in the standard error associated with differences in sample sizes becomes more pronounced for scores at greater distances from the mean raw score for the examination. In Figure 1, for example, the difference between the standard error for samples of size 15 and samples of size 100 is less than one point at the mean score of the examination, but increases to nearly two points at a score two standard deviations below the mean. The effect of sample size on the stability of the standard error of equating across raw score points is most striking for the samples of size 500,

which are included in the figures for reference. The flatness of these error curves, in contrast to the small sample curves, highlights the effect observed in the small sample equatings.

A summary of the standard errors of equating for selected standard scores from the five subject area examinations are presented in Table 2. This table provides the estimates of the standard error of equating at the mean, and at points 0.5, 1.0, and 1.5 standard deviations above and below the mean. Differences in the standard errors across the five subject area tests examined are evident in this table. Examination form 26 evidenced the smallest standard errors across all sample sizes examined, and examination form 34 evidenced the largest standard errors.

Insert Table 2 about here

For the breadth of scores summarized in Table 2, all of the standard errors were below 0.5 for the samples of size 500, with the exception of the most extreme scores (+/- 1.5 SD) on examination form 34. For samples of size 100, all of the standard errors were below 1.0 on exam forms 1, 26, 28, and 39. Only exam form 26 presented standard errors consistently less than 1.0 for samples of size 50, although exam form 28 exceeded 1.0 only for scores 1.5 SD below the mean. At sample size 25, the extreme

values of the standard error for exam form 26 were nearly 1.5 points. In contrast, exam forms 1 and 39 evidenced standard errors of nearly 2 points for samples of size 25 and scores 1.5 SD below the mean. For exam form 34, the standard errors of equating for samples of size 25 were approximately 2.5 points for scores 1.5 SD above or below the mean.

Graphical representations for the statistical bias in equating for each sample size are presented in Figures 6 through 10. The most striking characteristic of these figures is the apparently trivial magnitude of the equating bias, except for the biases obtained under the smallest sample size examined. However, even for samples of size 15, the magnitude of the bias is small except for extreme values of raw score. The pattern of the statistical bias, when bias is evident, is consistent across test forms -- the bias is negative for scores below the mean raw score and is positive for scores above the mean. In congruence with the results obtained for standard errors of equating, the smallest bias was obtained for exam form 26 (Figure 7), and the largest was obtained for exam form 34 (Figure 9).

Insert Figures 6 - 10 about here

A summary of the statistical bias for selected standard scores from the five subject area examinations are presented in Table 3.

This table provides the empirical estimates of the bias of equating at the mean, and at points 0.5, 1.0, and 1.5 standard deviations above and below the mean. The values of the statistical bias for raw scores within 1.5 SD of the mean are all less than one-third of a point, and 89% of the estimates of bias presented in Table 3 are less than one-tenth of a point. These results suggest that statistical bias related to small sample sizes for scores within 1.5 SD of the mean is not a principal concern.

Insert Table 3 about here

Because of the consistency of differences observed across subject area tests in both the standard error of equating and statistical bias (i.e., best results obtained for exam form 26, and worst for exam form 34), the correlations between the anchor tests and total tests were examined. These correlations are presented in Table 4. The highest correlations are evident for subject area test 26 ($r = 0.98$ for both forms of this test) and the lowest are seen for subject area test 34 ($r = 0.91$ for form A and $r = 0.92$ for form B). As suggested by the work of Budescu (1985) and Woodruff (1989), the magnitudes of these correlations are related to the relative success of the test equatings.

Insert Table 4 about here

The general question of whether linear test equating is recommended for small samples of examinees can certainly not be answered globally. The consistent finding in this study that statistical bias in equating is minimal with small samples supports its use. However, consideration must be given to what magnitude of standard error of equating is viewed as acceptable. The use of the test scores must be considered in this context. For applications in which a pass/fail decision is rendered based upon the total test score, focus should be on the standard errors of equating at and near the cutting score. If the cutting score is near the mean score for the test, where standard errors are smallest, equating can probably be supported for small samples. When cutting scores are not in the proximity of the mean, the magnitudes of the standard errors in the region of the cutting score may make the small sample results less tolerable. Similarly, for testing applications that make substantive use of the full range of test scores, rather than yielding a pass/fail decision, consideration must be given to the standard errors across the breadth of the score scale.

The importance of consideration of the standard errors of equating should not be underestimated. Formulas for estimating standard errors of linear equating (Angoff, 1984; Kolen, 1985; Lord, 1950) are intended for large sample situations (Peterson, Kolen, & Hoover, 1989). Bootstrap estimates of standard errors (Efron, 1982) are likely to provide more accurate estimates of standard errors for small samples. In this regard, Kolen (1985) suggested that formulas for standard errors of equating are accurate for samples only as small as 250. The examination of bootstrap estimates of the standard errors of equating should certainly be a routine component of applied equating operations.

An important limitation to this study is that the five subject area tests examined in this research were anchored in the most efficient way -- about 50% of the test items were anchors and the anchors covered the breadth of content coverage (Budescu, 1985). The research of Klein and Jarjoura (1985), and Peterson, Marco and Stewart (1982) suggest that the success of equating may be severely affected for tests in which anchors are considerably shorter or which do not match the total test content. Thus, a natural extension of the research reported in this paper is an empirical examination of the combined influence of small sample size and short or unrepresentative anchor tests.

An additional area of inquiry suggested by this research is that of evaluative indices of the accuracy of test equating. Although

we found that the value of the correlation between the anchor test and the total test was related to the success of equating, this population value is not known in actual field applications of test equating. Further, for all five examinations included in this research, some samples yielded equations that were identical to the equation computed from the pseudo-populations while other samples yielded equations that were very different from the population equation. Research on the sample characteristics that are associated with the accuracy of equatings should be undertaken. In this regard, Beard (personal communication, 1990) suggested that the degree of similarity in the shapes of the score distributions for the two test forms is a potentially useful index for predicting the success of test equating. Empirical research on the effectiveness of an index based on this similarity, as well as indices based on the additional similarities suggested by Jaeger (1980) would provide a useful tool for informing decisions about the advisability of linear equating in small samples.

References

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service. (Reprint of chapter in R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 1971.)

Angoff, W. H. & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. Journal of Educational Measurement, 23, 327-345.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22, 13-20.

Cook, L. L. & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Jaeger, R. M. (1980, April). Some exploratory indices for selection of a test equating method. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Klein, L. W. & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. Journal of Educational Measurement, 22, 197-206.

Kolen, M. J. (1985). Standard errors of Tucker equating. Applied Psychological Measurement, 9, 209-223.

Kolen, M. J. & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of General Educational Development. Journal of Educational Measurement, 19, 279-293.

Lord, F. M. (1950). Notes on comparable scales for test scores (RB-50-48). Princeton, NJ: Educational Testing Service.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). New York: American Council on Education

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Woodruff, D. J. (1989). A comparison of three linear equating methods for the common-item nonequivalent-populations design. Applied Psychological Measurement, 13, 257-261.

Table 1
Descriptive data on pseudo-populations.

Exam Number	Content Area	Test Form	N of Examinees	Total Items	Anchor Items	Raw Score Mean	Raw Score Std Dev
01	Art Education (K-12)	A	233	105	68	70.62	9.10
		B	270	109	68	73.47	8.35
26	Mathematics (9-12)	A	633	99	68	55.78	14.34
		B	519	99	68	53.46	13.54
28	Music Education (K-12)	A	259	100	58	71.42	8.35
		B	352	99	58	71.24	9.80
34	Early Childhood (K-3)	A	1311	146	69	98.24	12.04
		B	1105	144	69	92.82	12.73
39	Spanish (K-12)	A	242	91	47	68.68	10.51
		B	274	91	47	65.99	9.99

Table 2

Standard Error of Equating at Selected Score Points.

Exam Number	Sample Size	Score Point						
		-1.5 SD	-1.0 SD	-0.5 SD	Mean	+0.5 SD	+1.0 SD	+1.5 SD
1	15	2.23	1.76	1.36	1.09	1.08	1.32	1.72
	25	1.62	1.29	1.00	0.83	0.85	1.04	1.33
	50	1.13	0.90	0.70	0.57	0.58	0.71	0.92
	100	0.79	0.62	0.48	0.39	0.39	0.49	0.63
	500	0.37	0.29	0.22	0.18	0.18	0.22	0.29
26	15	2.10	1.67	1.30	1.06	1.04	1.26	1.61
	25	1.51	1.19	0.93	0.77	0.79	0.96	1.24
	50	1.01	0.81	0.65	0.54	0.54	0.65	0.81
	100	0.72	0.58	0.47	0.39	0.38	0.44	0.55
	500	0.32	0.26	0.21	0.17	0.16	0.19	0.24
28	15	2.65	2.04	1.51	1.14	1.14	1.49	2.03
	25	1.81	1.42	1.09	0.88	0.89	1.12	1.46
	50	1.21	0.95	0.73	0.59	0.61	0.76	0.99
	100	0.86	0.68	0.52	0.43	0.44	0.54	0.70
	500	0.39	0.30	0.23	0.19	0.19	0.24	0.31
34	15	4.19	3.31	2.54	1.98	1.86	2.24	2.93
	25	3.10	2.46	1.89	1.50	1.43	1.73	2.25
	50	2.10	1.66	1.28	1.01	0.97	1.17	1.52
	100	1.41	1.13	0.88	0.73	0.71	0.85	1.08
	500	0.63	0.50	0.39	0.31	0.30	0.36	0.46
39	15	2.76	2.14	1.58	1.17	1.10	1.42	1.94
	25	2.08	1.62	1.21	0.90	0.82	1.03	1.40
	50	1.45	1.13	0.84	0.62	0.58	0.73	1.00
	100	0.98	0.76	0.57	0.44	0.41	0.51	0.68
	500	0.44	0.34	0.25	0.19	0.18	0.23	0.31

Table 3

Statistical Bias of Equating at Selected Score Points.

Exam Number	Sample Size	Score Point						
		-1.5 SD	-1.0 SD	-0.5 SD	Mean	+0.5 SD	+1.0 SD	+1.5 SD
1	15	-0.18	-0.13	-0.08	-0.04	0.00	0.05	0.10
	25	-0.05	-0.03	-0.01	0.00	0.02	0.04	0.06
	50	-0.03	-0.02	-0.01	0.00	0.01	0.02	0.04
	100	-0.05	-0.03	-0.01	-0.00	0.01	0.02	0.04
	500	-0.04	-0.03	-0.02	-0.01	-0.00	0.00	0.01
26	15	0.03	0.03	0.03	0.03	0.03	0.04	0.04
	[illegible]	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	50	-0.03	-0.02	-0.01	-0.00	0.01	0.02	0.03
	100	0.01	0.01	0.01	0.00	0.00	0.00	-0.00
	500	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01
28	15	-0.14	-0.09	-0.04	-0.00	0.04	0.09	0.14
	25	-0.06	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
	50	-0.09	-0.07	-0.04	-0.02	-0.00	0.02	0.04
	100	-0.04	-0.04	-0.03	-0.03	-0.02	-0.02	-0.01
	500	-0.02	-0.03	-0.03	-0.03	-0.03	-0.04	-0.04
34	15	-0.26	-0.18	-0.10	-0.02	0.05	0.13	0.21
	25	-0.15	-0.09	-0.02	0.04	0.11	0.18	0.25
	50	-0.04	-0.02	-0.00	0.01	0.03	0.06	0.08
	100	0.04	0.04	0.04	0.04	0.04	0.03	0.03
	500	0.00	0.01	0.02	0.02	0.03	0.04	0.05
39	15	-0.07	-0.04	-0.01	0.01	0.05	0.08	0.11
	25	-0.19	-0.16	-0.12	-0.09	-0.05	-0.02	0.01
	50	-0.07	-0.05	-0.04	-0.02	-0.01	0.00	0.02
	100	-0.01	-0.02	-0.02	-0.02	-0.02	-0.03	-0.03
	500	0.00	-0.00	-0.01	-0.02	-0.03	-0.04	-0.04

Table 4

Correlations Between Anchor Test and Total Test.

Exam Number	Form A Correlation	Form B Correlation
01	0.95	0.94
26	0.98	0.98
28	0.93	0.95
34	0.91	0.92
39	0.96	0.96

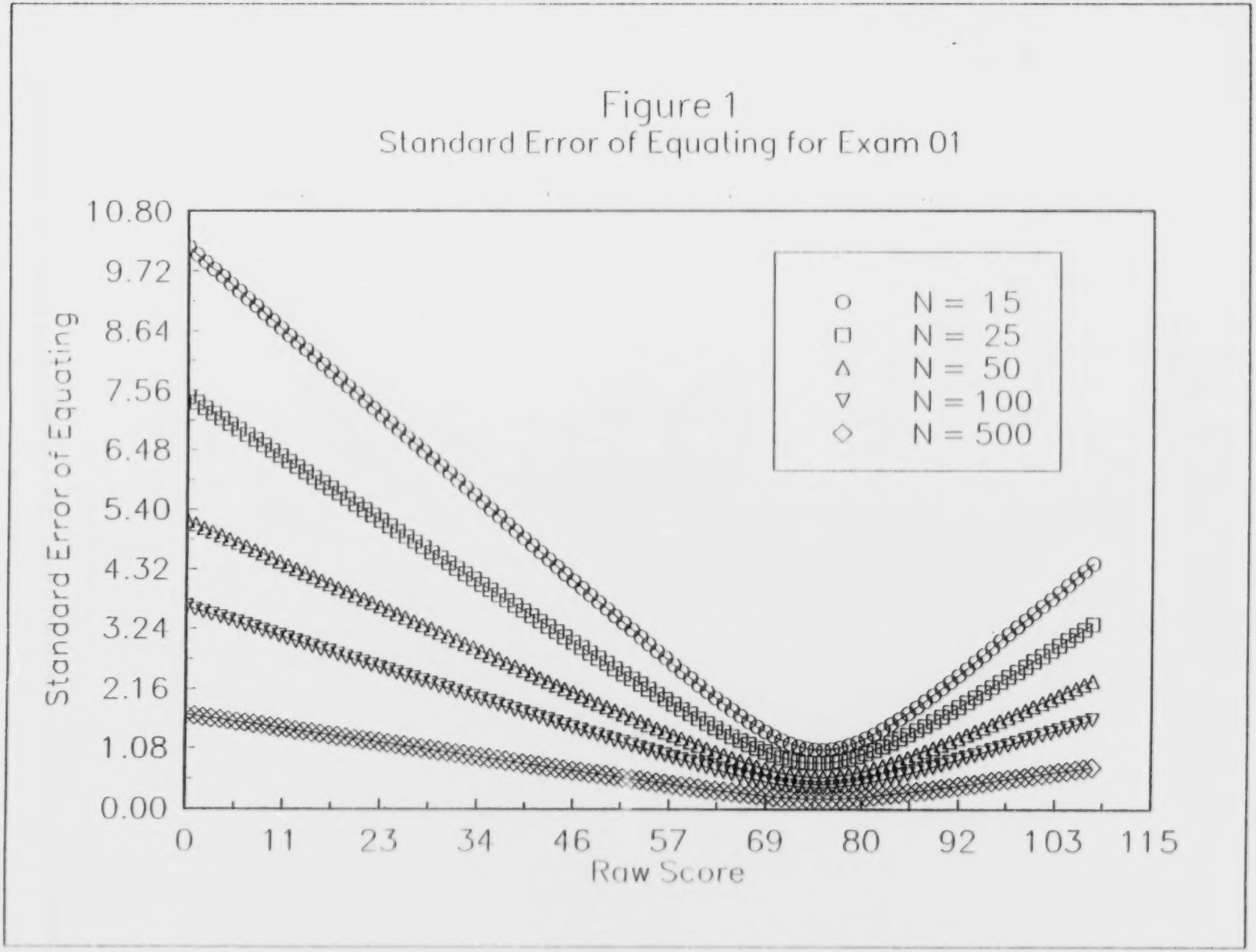
Figure 1
Standard Error of Equating for Exam 01
Standard Error of Equating
10.80
9.72
8.64
7.56
6.48
5.40
4.32
3.24
2.16
1.08
0.00
0
11
23
34
46
57
69
80
92
103
115
Raw Score
N = 15
N = 25
N = 50
N = 100
N = 500

Figure 2
Standard Error of Equating for Exam 26

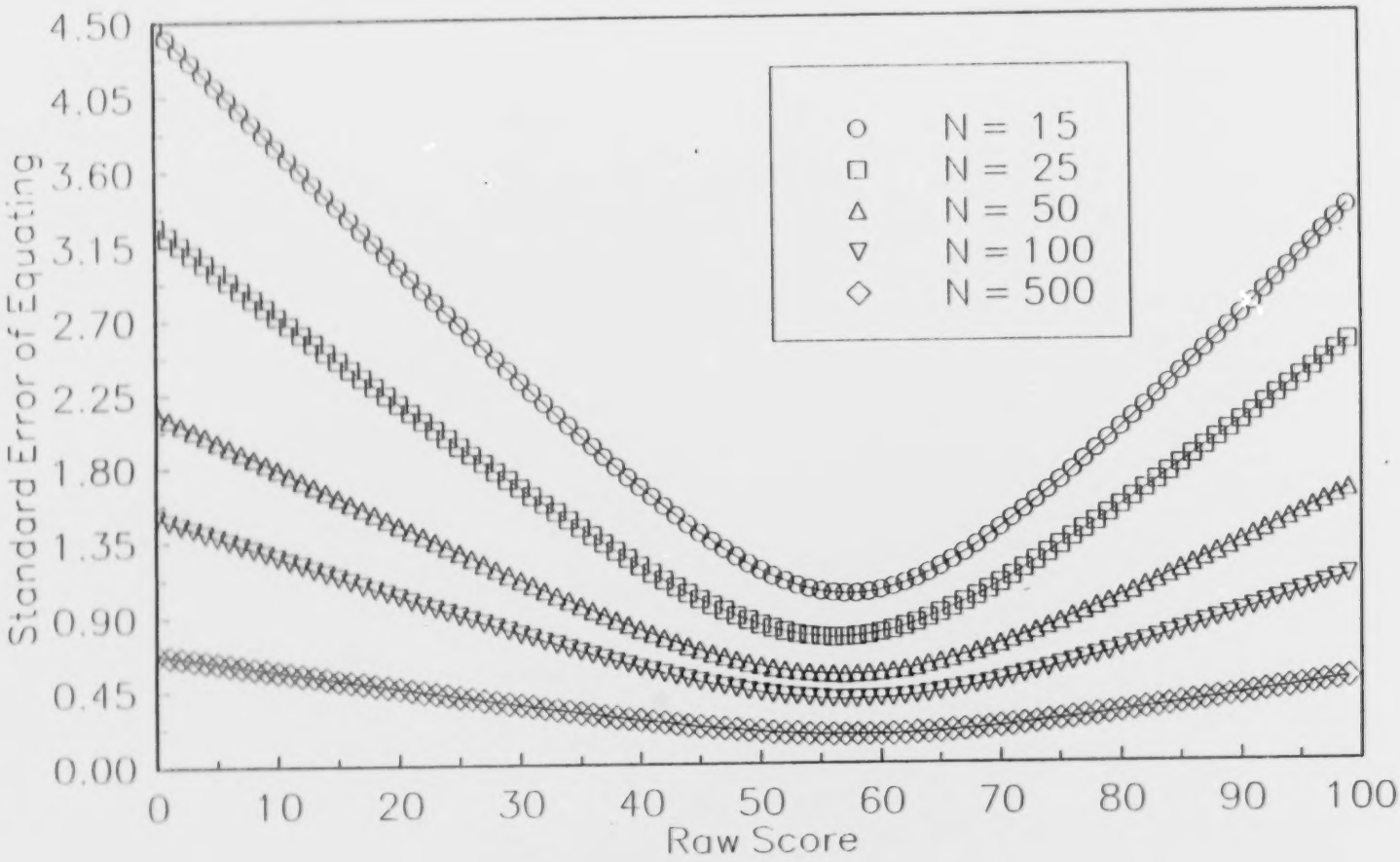

Figure 3
Standard Error of Equating for Exam 28

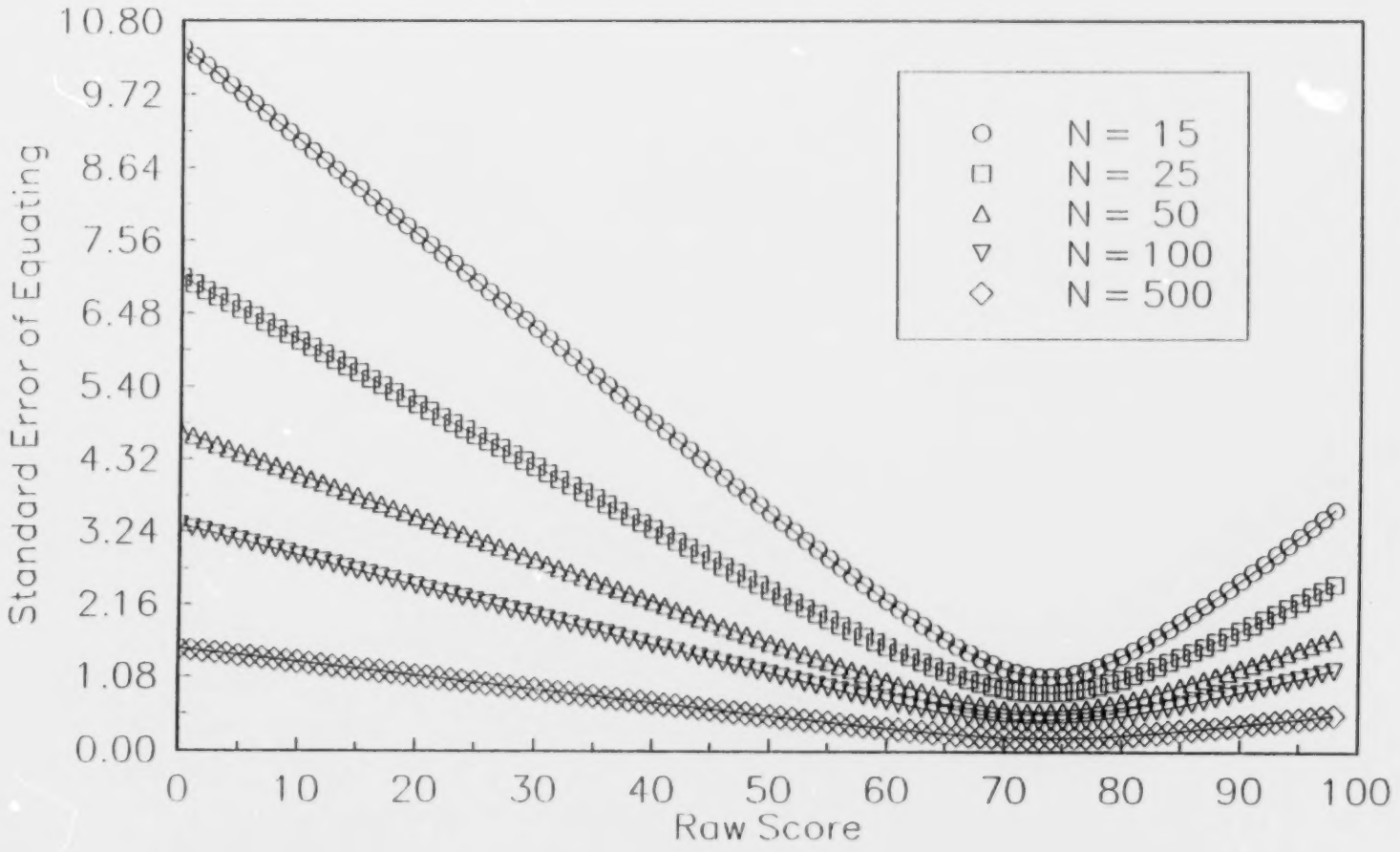

Figure 4
Standard Error of Equating for Exam 34

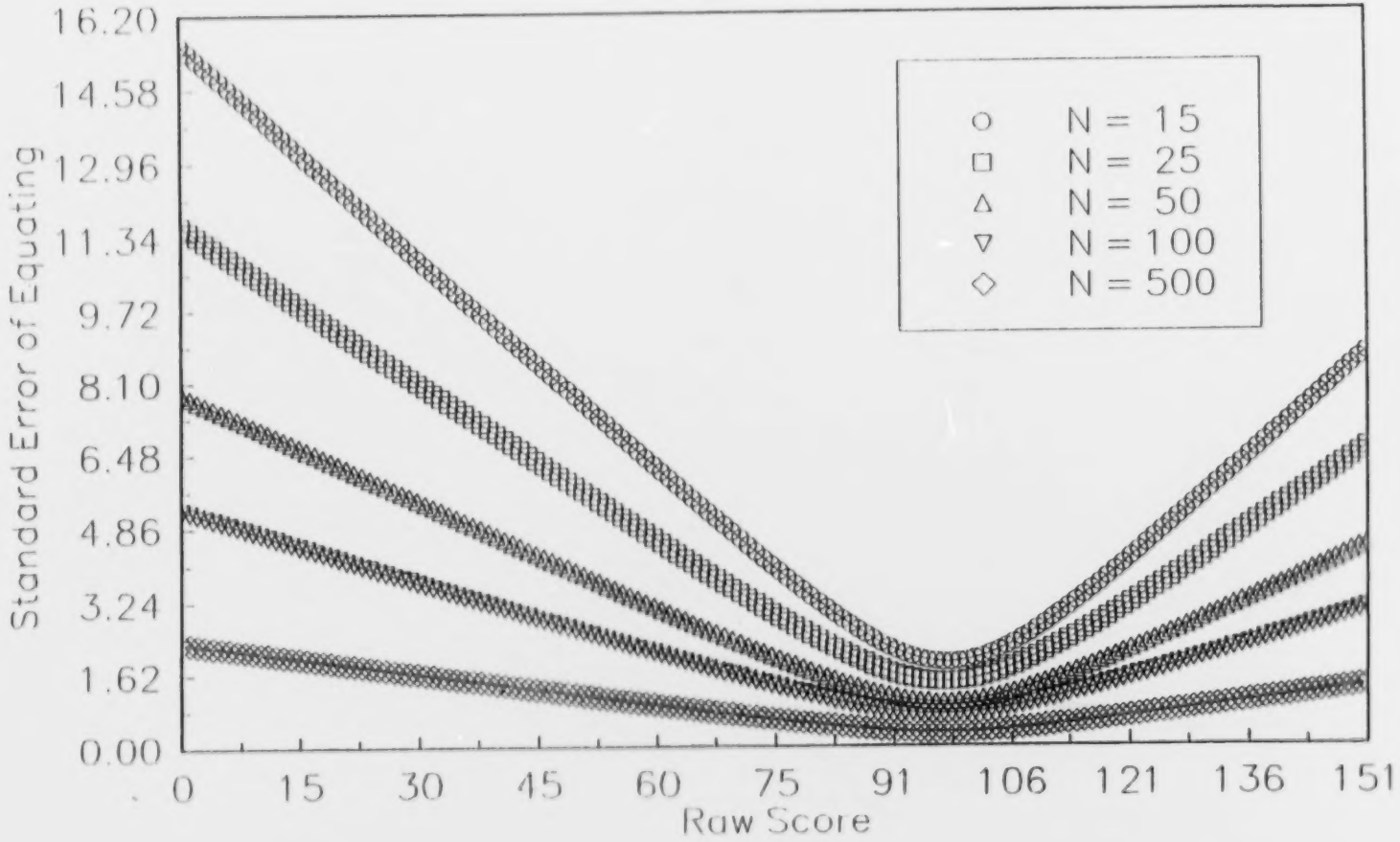

Figure 5
Standard Error of Equating for Exam 39

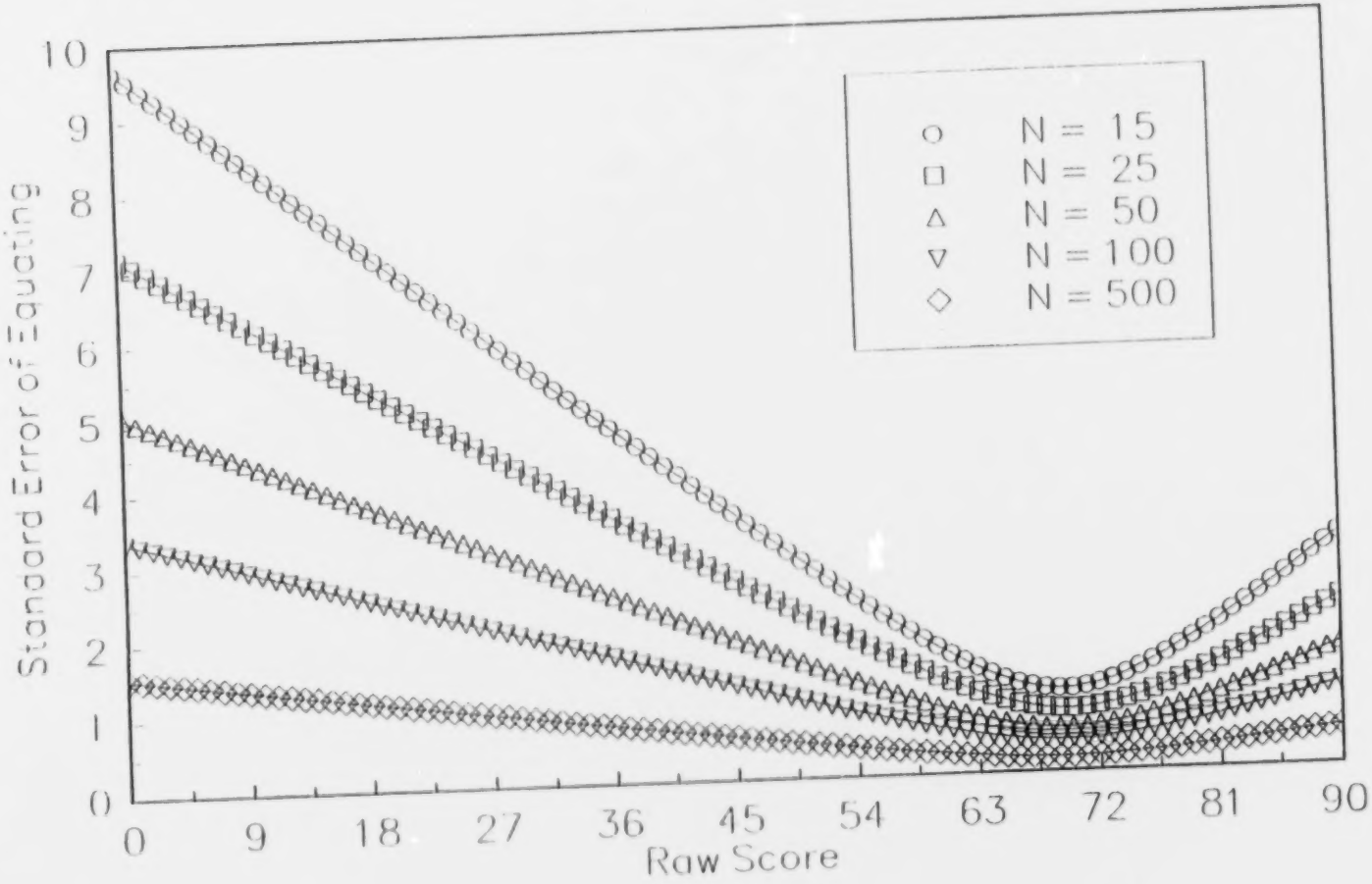

Figure 6
Equating Bias for Exam 01

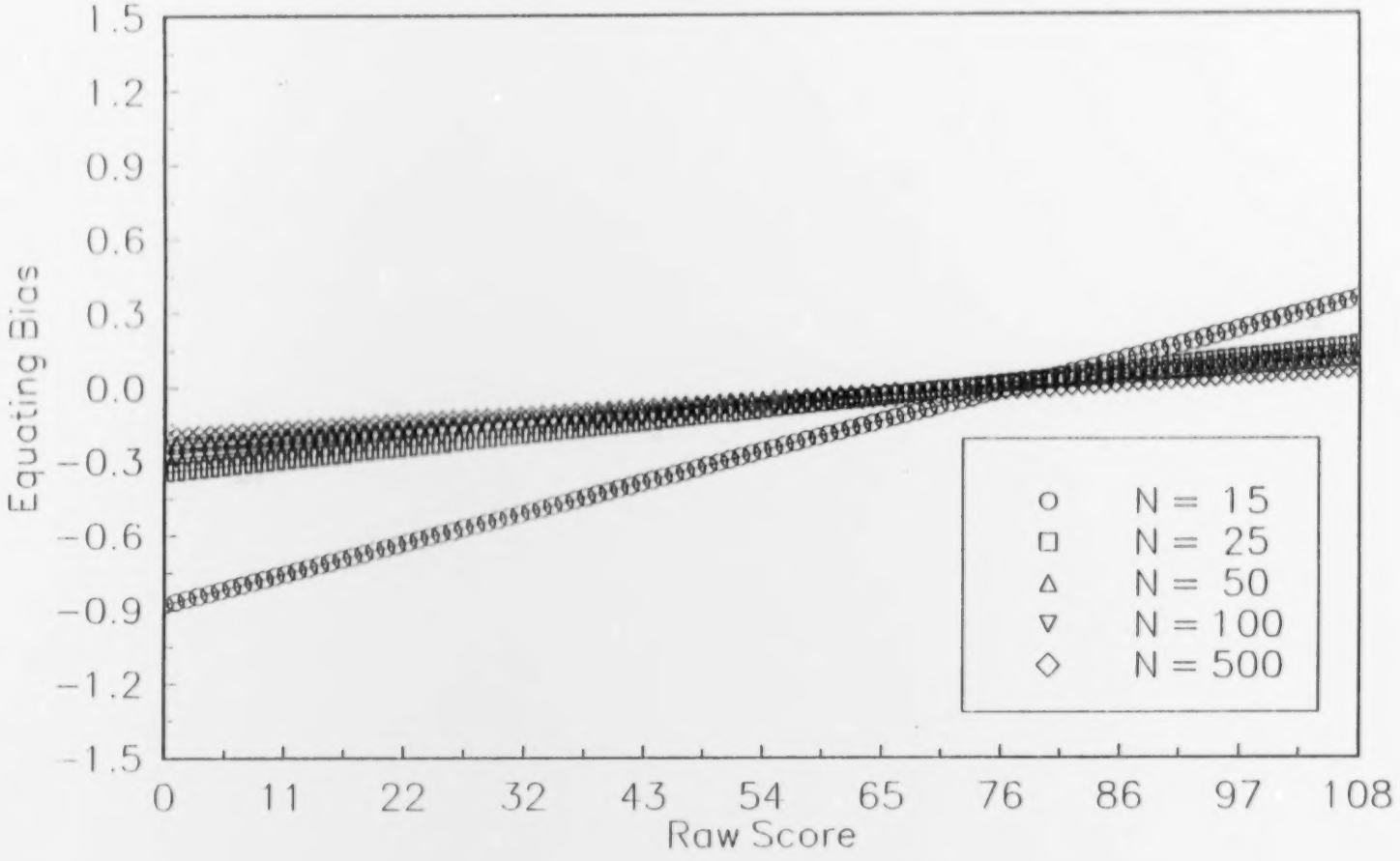

Figure 7
Equating Bias for Exam 26

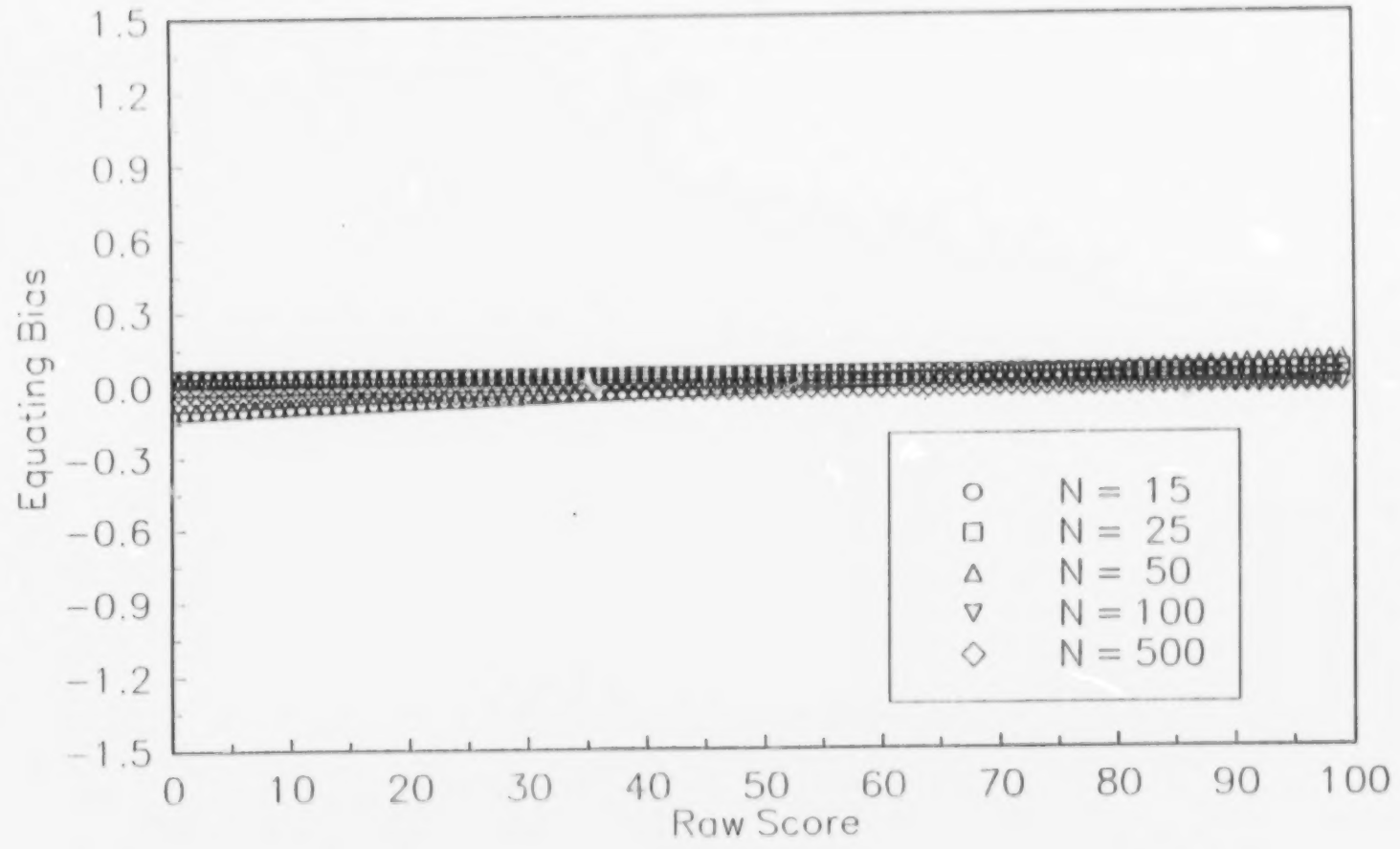

Figure 8
Equating Bias for Exam 28

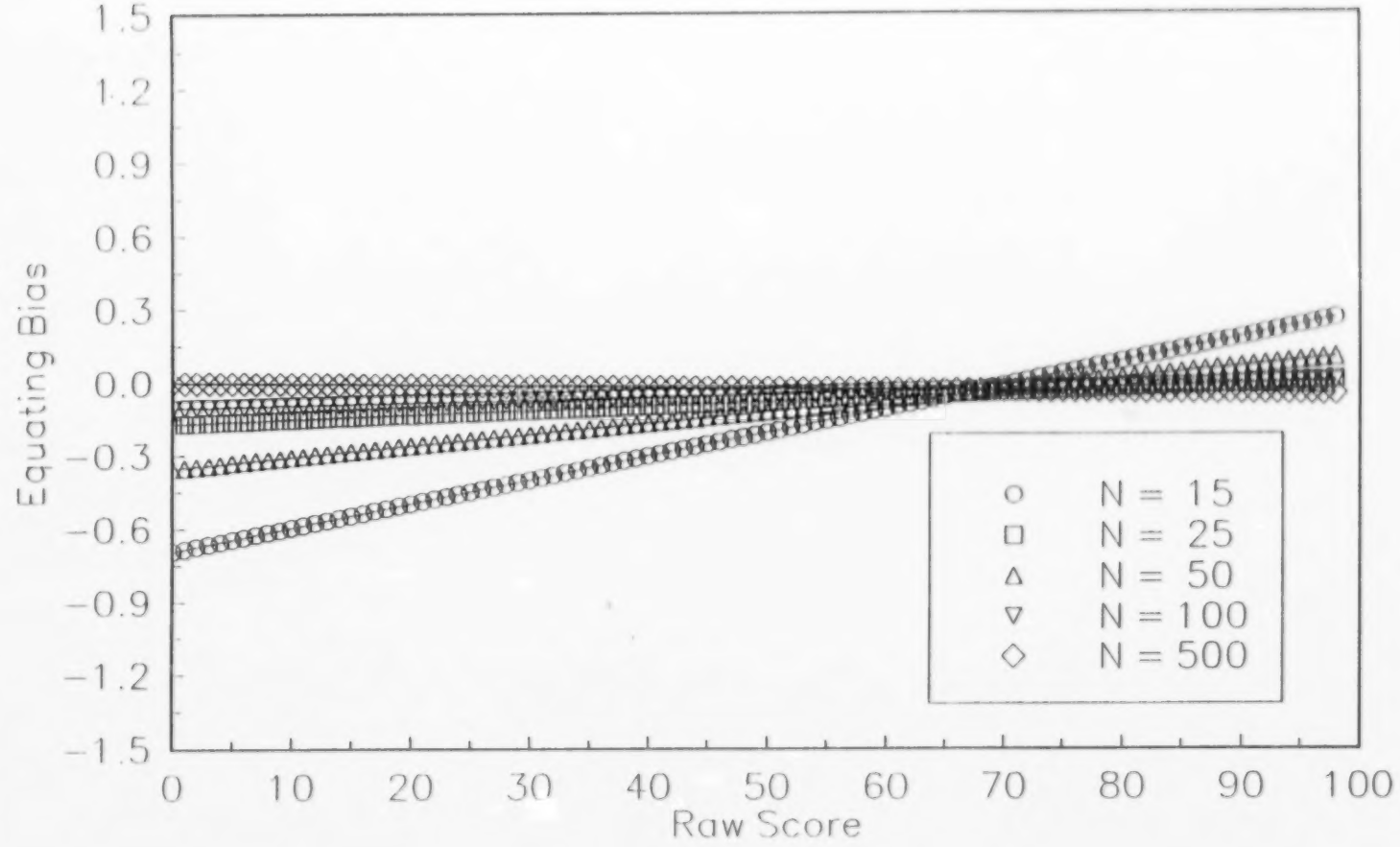

Figure 9
Equating Bias for Exam 34

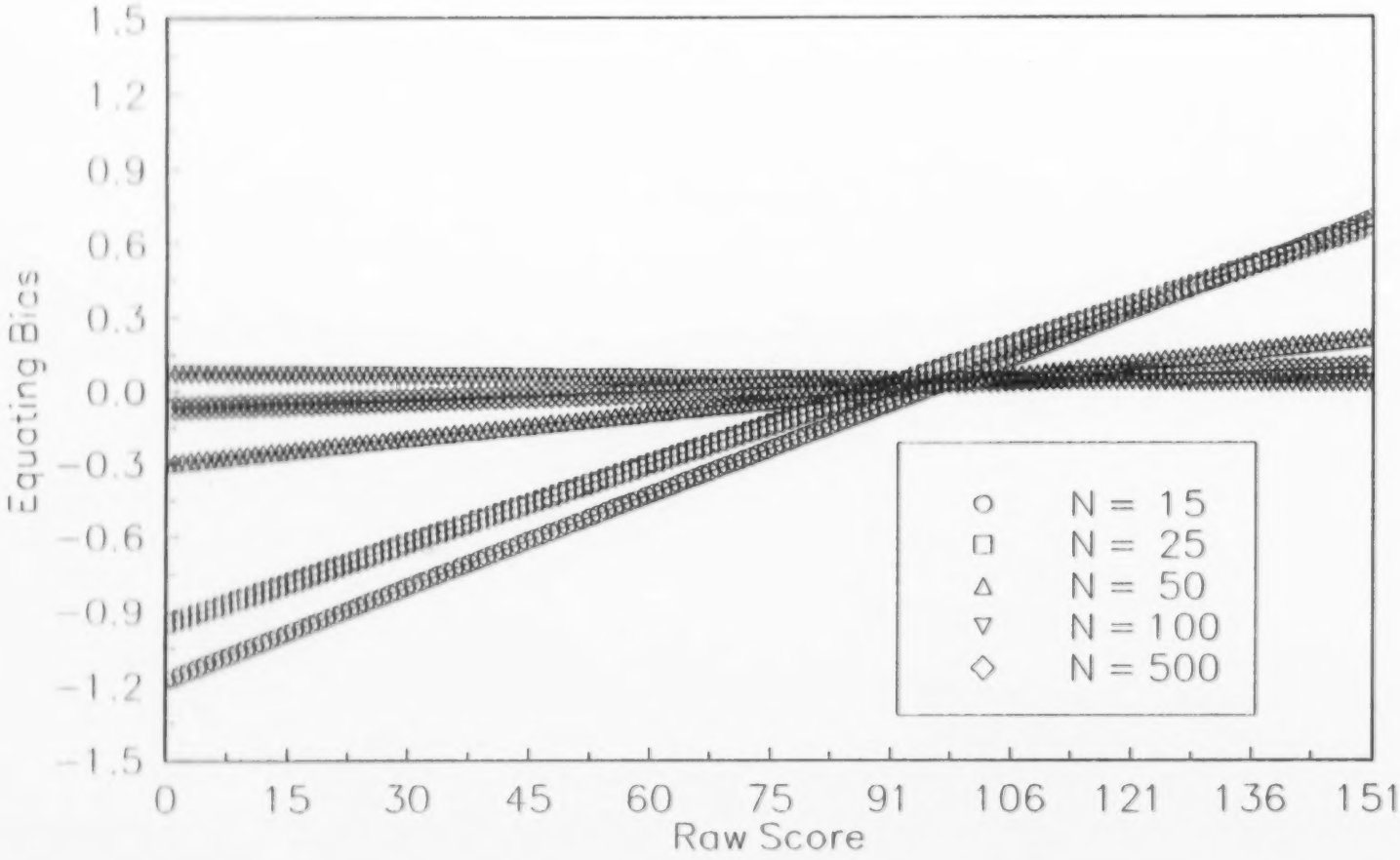

Figure 10
Equating Bias for Exam 39